**HPAM 9000 Directed Research in Health Policy and Management**

**Spring 2024**

**CRN:**

**Hours:** 3 Hours

**Course Title:** Using Python to Integrate Diverse Data for Machine Learning

**Instructor:**

**Meeting Time:** Bi-weekly 60 minutes

**Course Overview:**

This course is designed to provide in-depth training in using Python to prepare and analyze complex, multidisciplinary datasets for artificial intelligence (AI) and machine learning applications. It covers various topics, from data cleaning and preprocessing to advanced feature engineering and data structure optimization. The course emphasizes practical, hands-on experience with Python, focusing on real-world datasets that include environmental, ecological, biological, land use, and demographic data and their integration using spatial and temporal attributes.

**Course Goal:**

The primary goal of this course is to equip students with the necessary skills to use Python to prepare and analyze diverse datasets for AI applications. This includes understanding and implementing best practices in data design, structure, and integrity to enhance the accuracy and efficiency of AI models.

**Course Learning Objectives:**

1. Proficiency in Data Cleaning and Preprocessing Techniques:
   1. Gain expertise in using Python for cleaning and preprocessing multidisciplinary datasets, focusing on handling missing data, data normalization, and ensuring proper data structure.
2. Develop Skills in Feature Engineering:
   1. Learn to create, select, and transform data features using Python, enhancing their relevance and effectiveness for machine learning models.
3. Understand and Implement Data Integration Methods:
   1. Acquire skills in integrating diverse data types using Python techniques.
4. Apply Best Practices in Data Structure and Design:
   1. Understand the principles of data design and structure that contribute to the accuracy and performance of AI models and apply these principles in practical settings.
5. Conduct Comprehensive Data Analysis and Description:
   1. Develop the ability to use Python for advanced data analysis, including assessing distributions and providing statistical descriptions of complex datasets.
6. Create Reusable Python Scripts for Data Preparation:
   1. Learn to develop and document Python scripts that are modular and adaptable for various AI/machine learning projects, enhancing efficiency in data preparation tasks.
7. Evaluate Data Quality and Model Readiness:
   1. Gain proficiency in assessing datasets' quality, consistency, and readiness for AI modeling, ensuring data integrity and suitability for machine learning applications.

**Selected Doctor of Public Health Program Competencies Addressed:**

1. Data & Analysis:
   1. Competency 1: Explain qualitative, quantitative, mixed methods, policy analysis research, and evaluation methods.
   2. Competency 3: Explain the use and limitations of various data sources in assessing, monitoring, and evaluating policies and programs.
2. Policy & Programs:
   1. Competency 14: Design system-level interventions to address a public health issue using multidisciplinary data.
3. Additional Competencies in Health Management and Policy:
   1. Competency 3: Develop proficiency in data science and analytics to address research and practical questions related to population health.

**Course Assignments:**

1. Python Data Cleaning and Preprocessing Project:
   1. Task: Clean and preprocess a complex multidisciplinary dataset using Python, focusing on missing data, proper structure, and spatial/temporal data joins.
      1. Deliverable: A report documenting the preprocessing steps, challenges, solutions, and the Python code.
2. Data Analysis and Description Exercise:
   1. Task: Use Python to perform statistical analysis and describe the dataset, exploring relationships between various data types.
      1. Deliverable: Python script for analysis and a report on statistical findings and their implications for AI applications.
3. Feature Engineering and Selection Assignment:
   1. Task: Apply feature engineering to enhance a dataset and select relevant features for machine learning models.
      1. Deliverable: A report on the feature engineering process, feature selection rationale, and impact on model performance.
4. Practical Exercise on Data Normalization and Transformation:
   1. Task: Normalize and transform a provided dataset using Python for machine learning readiness.
      1. Deliverable: Python code and a report on the transformation process.
5. Development of Reusable Python Scripts:
   1. Task: Create adaptable Python scripts for common data preparation tasks in AI/machine learning projects.
      1. Deliverable: A collection of Python scripts with usage and modification documentation.
6. Dataset Structure and Integrity Assessment:
   1. Task: Develop a Python-based methodology to assess dataset integrity and structure for machine learning.
      1. Deliverable: Python script for dataset assessment and a methodology report.
7. AI/Machine Learning Model Preparation with Emphasis on Data Structure:
   1. Task: Prepare a dataset for a specific AI application, including all stages of data preparation, emphasizing data structure optimization.
      1. Deliverable: A comprehensive report detailing the process, Python code, challenges, and how the dataset structure impacts model accuracy.